考古发掘资料图数据库的语义关联构建研究*

■ 高劲松　韩牧哲
华中师范大学信息管理学院　武汉 430079

摘　要：[目的/意义]针对原始资料整理中存在的问题,提出一种可实现考古发掘资料数据化转换和语义关联的方法,帮助考古学工作者规避低效流程。[方法/过程]首先,结合实例对人文学科原始资料的特征进行解析,设计原始资料数据化转换的过程和方法;其次,选取新疆和静察吾呼墓地的考古发掘资料为实证数据来源,构建考古发掘资料图数据库;最后,以文物间的共存关系为例,实现考古发掘资料图数据库的语义关联构建。[结果/结论]考古发掘资料图数据库及其语义关联的构建,为考古发掘资料的数据化转换提供了新的方法和思路,在数字人文领域有推广价值和实际意义。

关键词：考古发掘资料　图数据库　语义关联　察吾呼墓地
分类号：G255
DOI：10.13266/j.issn.0252-3116.2021.09.012

1　引言

考古学在人文学科中属于史学大类,是较早采用数字人文技术作为研究手段的学科之一。在数字人文实践的6个主要方向中,“历史学方面的基于GIS的历史地理可视化”和“考古学方面的图像分析、色彩还原和数字重建”都是与考古学密切相关的数字人文方向[1]。在基于DH Commons所做的英国数字人文项目统计中,有542个标注了研究领域的项目,其中考古学被标注了88次,并认为“历史研究(含考古)、语言文学研究、图书馆/信息和博物馆研究是英国数字人文项目最重要的研究领域”[2]。

目前,相较于考古学领域,直接与图书馆、档案馆、博物馆(Library, Archive and Museum, LAM)机构结合的文物和文化遗产领域对于数字人文及面向人文学科的知识服务的学术敏感性更强,这得益于我国在过去20多年间已经基本完成的LAM机构馆藏资源大规模数字化转换进程[3]。然而,在考古发掘工作过程中,会持续产出零散且种类丰富的考古发掘资料,其通常不具备直接进行结构化组织和数据化应用的条件,这使得新发现的资料难以融入现有的知识组织体系,进而影响到考古学数字人文研究的整体进程。因此,探索适用于考古发掘资料直接录入与整理的数据组织和语义关联构建方法有重要的现实意义。

2　相关研究

目前,国内外在考古信息资源的组织、存储、管理与应用方面均取得了相应的实践成果,较具代表性的包括英国ARCH资助的STAR项目[4-5]和雷丁大学的IADB数据库[6]、德国的Archeo-Info系统[7]、美国芝加哥大学的OCHRE系统[8]以及中国社会科学院考古研究所和清华大学合作研发的E-Arch系统[9]等。上述项目所包含的信息资源以经过整理的数字化资源为基础,但对于更原始的考古发掘资料而言,纸质档案仍然是国内外很多机构主要的保存方案。大量考古发掘资料会按照传统方法以实物和数字化副本形态分布存储在不同的机构中,这种资源存储方式会造成非常严重的知识揭示与分享障碍,蕴含着丰富知识的发掘资料在有序性、开放性、安全性乃至研究价值上都会因此大打折扣。其中,在考古发掘资料的记录、整理和数据化过程中所面临的效率和技术难题是不容忽视的。

综合数字人文其他领域的成果来看,有关技术和

* 本文系中央高校基本科研业务费自由探索项目“面向用户的文物信息资源知识服务研究”(项目编号:CCNU20A06025)和国家社会科学基金重大项目“新时代我国文献信息资源保障体系重构研究”(项目编号:19ZDA345)研究成果之一。
作者简介:高劲松(ORCID:0000-0003-0022-5923),教授,博士生导师,E-mail:jsgao@mail.ccnu.edu.cn;韩牧哲(ORCID:0000-0002-8474-4570),博士研究生。
收稿日期:2020-11-30　修回日期:2021-02-06　本文起止页码:105-116　本文责任编辑:徐健

方法的研究主要集中于数据库建设和知识可视化，以及在此基础上的学科服务平台开发，其热点话题涵盖数字人文领域的关联数据发布[10]、文本挖掘[11]、元数据组织[12]、本体建模[13]和知识图谱开发[14]等。上述研究涉及实证部分的数据源通常来自LAM馆藏体系中已有的结构化和半结构化的数字资源，通常具有较为确定的框架结构和可参照的元数据控制方案，使研究者可以根据信息需求完成知识抽取和知识表示，以及进一步的关系型数据库的构建、知识图谱的开发和后续的知识服务工作，对于异构的外部数据库也可以通过知识融合方法实现数据互联。

从过程上分析，上述研究起点多为知识组织环节，其面对的数据通常是经过数字化和文本化梳理的二次或三次情报，而本文的研究对象——考古发掘资料，则属于更基础的一次情报，对应的是从各个场景中采集到的原始资料，在考古学以外的其他人文学科中也普遍存在。在数字人文研究的前期，几乎都要经历对LAM机构馆藏资源进行数字化(Digitalization)转换的阶段，一般做法是将结构化水平很低的资源进行结构化整理后存入关系型数据库，该阶段需要借助大量的人力参与，尤其是需要人文学者从事大量低水平且繁杂的资料搜集和整理工作[15]。从宏观进程来看，国内对大规模馆藏资源的数字化转换阶段已经告一段落，但在人文学科领域，与考古发掘资料类似的新的原始资料仍在不断产生，各领域人文学者们以传统的低效方法进行资料采集和整理的现象依然普遍存在，而当前数字人文领域的技术研究对于这些零星产生于各学科一线、不适合或无法直接纳入关系型数据库的原始资料并无妥善的解决方案。因此，笔者以人文学者的现实需求导向，面向数字人文研究过程，构建一种人文学科的原始资料整理模型，以考古发掘资料为对象，提出能够实现原始资料数据化(Datalization)转换的数据库构建方法，以及在此基础上实现语义关联，以帮助考古工作者及其他领域人文学者规避工作中的低效流程，为进一步的语义化知识服务做好数据准备。

3 原始资料的分析与整理

3.1 原始资料的特征解析

原始资料(Primary source)是指包含了原始信息的文献、实物、现象和其他事物，从中直接获取的原始信息往往结构不统一，资料分布更为碎片化。一份资料所承载的是否是未经过知识揭示的原始信息，是界定其是否为原始资料的标准。常见的LAM数据资源实例可归纳为“无结构数据”“半结构化数据”和“结构化数据”3类，无结构数据的表现为“文献、文物、物件自带的原始信息”，这些原始信息及其载体均可被视为原始资料[16]。在考古工作中，通过田野调查和考古发掘直接记录、采集、汇总、统计所得的文献、实物和数据资料均属于原始资料。换言之，考古学中的原始资料可细分为田野调查资料和考古发掘资料。在针对考古发掘资料的整理方法进行讨论之前，有必要对人文学科中原始资料的一般特征进行解析。

相较于自然科学和社会科学，人文学科知识元素之间的关联关系与发展演化规律更加隐晦，这使得对其知识框架整理的难度也有所加大，且对对象和过程的描述通常具有主观性。因此，在研究工作完成之前，很难对其中的大量内容和知识直接进行结构化描述与存储。经过人文学者的研究，通常会在外部结构化描述的基础上，依据从原始资料中抽取出的信息对其进行分类命名、内容描述和其他知识揭示，有效知识揭示所产生的知识成果是原始资料具备进一步数据化转换的条件，即资料可以通过相应的数据进行描述、表达、存储和应用。职能上面向公众服务的LAM机构的馆藏结构中虽然包含了一定比例的原始资料，但其中大部分馆藏资料都有研究基础，为方便区分，笔者将有研究基础、具备数据化转换条件的资料称为馆藏资料。

人文学科中的原始资料在各方面的特征都与馆藏资料存在差异，详见表1。在研究的不同阶段，二者所面对的资料实体可能会发生重叠。总的来说，随着研究的深入，对资料解析程度的提升，原始资料会逐渐向数据化条件完善的馆藏资料转变。

表1 原始资料和馆藏资料(LAM机构)特征对比

对比项	原始资料	馆藏资料
资料来源	观察、田野调查、社会调查等	研究、整理和创作
资料内容	实物、现象、过程、概念及其衍生物	文献及与文献内容相关的实物
命名规则	缺乏参照，不确定或不完全确定	可参照，较为确定
分类规则	部分明确	明确
结构化效果	较差	好
数据规模	中小规模	大规模
产出模式	零散、持续且不规律	集中、单次或定期
操作人员	人文学者	人文学者和知识管理人员
研究基础	通常较差或无研究基础(新材料)	较好
使用目的	研究过程、资料保存	公共教育、研究服务、商业用途等
共享范围	业内或私密	公共领域或有条件开放
共享形式	非正式	正式

笔者以 1993 年出土于江苏邳州九女墩三号墩的一件铭文铜盘作为实例进行解析，解析细节如图 1 所示。“原始资料”栏的考古发掘资料来自 2002 年发表的考古发掘简报[17]，“馆藏资料”栏补充了 2019 年的铭文释读信息[18]。同一墓葬中出土的纹饰、器型相同的青铜盘共 5 件，且实例中的青铜盘是在清理过程中被发现底部刻有铭文。由于考古报告上并未对其做区分，故而原始资料栏中的外部结构化描述无法对这件铭文青铜盘给予准确的专属性命名。而随着研究过程的深入，铭文被释读之后，馆藏资料栏中基于研究内容对这件铭文青铜盘的结构化描述足以将其与共同出土的 4 件蟠蛇纹盘进行区分，此时才能将其作为结构化数据存储到关系型数据库中，为进一步的知识管理和知识服务工作做准备。

元素名称				原始资料	馆藏资料
基本信息	类别			青铜器	青铜器
	名称	器名		青铜盘	青铜盘
		异名		蟠蛇纹盘*	保媜盘**
	时代			春秋晚期	春秋晚期
	族属			徐国	徐国
	现藏地			Null	邳州市博物馆
器物内容	形制	度量		口径 22.6 厘米 高 10 厘米	口径 22.6 厘米 高 10 厘米
		器型 🔗		盘，口沿方折，颈略收，肩稍斜，弧腹，平底	盘，口沿方折，颈略收，肩稍斜，弧腹，平底
	纹饰 🔗			腹部饰两道绳纹，颈、腹部均饰细密、整齐的蟠蛇纹，腹下部饰两周三角纹，内填蟠蛇纹	腹部饰两道绳纹，颈、腹部均饰细密、整齐的蟠蛇纹，腹下部饰两周三角纹，内填蟠蛇纹
	铭文 🔗	位置		Unknown	器内底
		阴阳文			阴文
		字数			21 字，重 1
		行数			4 行
		释文	内容		叡句徐之孙、肤旨䣄之子保，保䞘公之妻媜，同铸用锗
			来源		马永强，程卫. 江苏邳州九女墩三号墩出土铜盘铭文考释[J]. 文物，2019(10):79-81.
出土信息	时间			1993 年	1993 年
	地点			江苏省邳州市戴庄镇	江苏省邳州市戴庄镇
	遗址			九女墩，三号墩	九女墩，三号墩
	类型			墓葬	墓葬
	器物编号			M3:43	M3:43
	来源			孔令远，陈永清. 江苏邳州市九女墩三号墩的发掘[J]. 考古，2002(5):19-30.	孔令远，陈永清. 江苏邳州市九女墩三号墩的发掘[J]. 考古，2002(5):19-30.

*按照青铜器类文物的命名规则，无铭文的青铜器通常依照特征纹饰+器型命名。

**带铭文的青铜器命名有“名从主人”原则，该铜盘为保与其妻媜两人所铸。

器型（照片）



纹饰（拓片）

铭文（拓片）

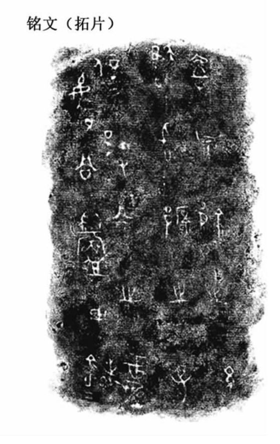


图 1　原始资料与馆藏资料的结构化描述效果对比

结合特征对比与实例解析可知，在相关研究取得一定成果之前，原始资料中的对象很难通过描述性命名和层级制分类进行定义。现有的结构化描述方法不能对原始资料中存在差异的对象进行辨识与区分，使得这些资料所转换的结构化信息即便被纳入现有关系型数据库和知识图谱中，也无助于进一步的学科研究和知识共享，还有可能导致语义模糊。但是，原始资料中的对象在知识网络中关系是相对稳定的，通过对对象在已有知识体系中已知关系的描述，可将原始资料转换为确定关系、开放命名、未定分类的中间态数据(Intermediate data)。由原始资料直接转化而来的中间态数据本身包含诸多未知或待定的属性，并不适合直接面向公众进行数据共享，但在功能上可以满足进一步人文研究的相应需求，帮助人文学者进行资料的录入、整理、统计与分析，促进原始资料向馆藏资料转换。此外，中间态数据以关系结构描述对象，本身具有相应的知识揭示功能，若能与目标领域的本体结构进行语义匹配，即可促使其标准化、规范化，作为外部数据库融合到现有的知识图谱中，直接实现原始资料的数据化转换。

3.2　原始资料整理方法

陈涛等将数字人文的研究进程划分为资源数字化转换、数字资源的文本建设和研究、对文本化资源的数据化和智慧化研究 3 个阶段，是对当前数字人文研究宏观进程的描述，并就此提出了宏观数字人文研究框架[16]。宏观研究框架中的馆藏资料将经过数字化转换和结构化描述，存储到关系型数据库中并转换为人文数据。我国在过去的 20 多年间已经基本完成了第一阶段的工作，目前正处于第二阶段颇具成效，向第三

阶段逐步迈进的过程。从微观上看,对于从各人文学科持续性产出的轻量化原始资料而言,在短时间内人文学者很难主动进行数字化转换和结构化描述,无法直接将其存储到关系型数据库中,按照宏观的数字人文研究框架,后续的文本化、数据化和智慧化也难以推进。尽管关系型数据库和语义知识图谱具有严格的元数据控制和本体结构,且在严谨性、标准化和长期保存与共享方面表现更为优秀,但对于原始资料的整理工作并非上佳之选。笔者认为有必要对人文学科中原始资料的处理方法和过程进行重构,促使其以较高的效率与宏观数字人文进程实现同步,较为可行的思路是对原始资料中方便进行结构化描述的内容进行外部结构化描述,复用现有关系型数据库的结构和元数据规范进行数字化和文本化;对于其他难以与现有关系型数据库进行规范、统一、标准化描述的内容,则以对象间的关系对其进行描述。

图数据库善于处理大量复杂、互连接、低结构化的数据,具有更强的数据兼容性,且对关联关系的表达更直观、处理更高效,其功能方面"更侧重于知识挖掘和计算,发现隐性知识并可视化,实现诸如提问式检索、时空展示等功能,推动人工智能环境下数字人文研究方法的创新"[16]。尽管图数据库简单易用,但是由于缺乏标准化的规范词表控制,不同图数据库之间难以互通,数据孤岛问题仍难以避免。不过,笔者对原始资料处理的直接目标是将其转换成中间态数据,这一思路是面向人文学者的研究过程而非面向公众的知识共享,图数据库在原始资料的录入、整理和存储方面更为适用。综合考虑多方面因素之后,笔者决定选择 Neo4j 进行考古发掘资料图数据库的构建。图数据库简单的"N-E"(Nodes & Edges,节点和边)和"K-V"(Keys & Values,键和值)结构可以包容大量中间态数据,同时其易用性也可以满足非专业人士的数据维护需求。

综上所述,宏观框架从整体上将数字人文研究进程分为 3 个阶段,考虑到微观形态上轻量化原始资料不断产出的过程,笔者引入图数据库并提出了可进一步与宏观框架关联的原始资料整理过程,如图 2 所示。将宏观研究框架的起点向前回溯:①在获取原始资料之后,由各个领域的人文学者接触并处理原始资料,依托人文研究方法对原始资料进行分类、整理、辨识、解析,使之转化为具有一定研究基础的馆藏资料,并与相关成果一同归档于典藏机构,这一过程本质上属于传统人文研究阶段,在宏观上属于数字人文研究的基础和前提,在数字人文兴起之前,这种传统的人文研究进程已经持续上百年,相应的人文研究成果也有着丰厚的积淀。②宏观数字人文研究框架的真正起点是对馆藏资料的数字化转换,自 21 世纪以来,在 LAM 等典藏机构先后开展的数字化建设和数字人文理念下推动了大规模的馆藏资料数字化,其主要工作是对各类馆藏资料做结构化描述和基于关系型数据库的存储,以元数据规范各类资料的著录信息,并将其作为索引指向各类数字化存储的信息资源、纸质文献和实物。③对馆藏数字化资源的文本化和文本分析、语义化和基于语义知识图谱的推理、共享是数字人文宏观研究在当前的主要方向,在此基础上开展的各项知识服务也是数字人文研究的目的。

但是,对于宏观研究框架而言,在传统人文研究和大规模数字化阶段所做的积淀是后续步骤推进的前提,这使得在微观上,近些年持续零星产出的人文学科原始资料在短时间内难以跟进宏观进程,因此,笔者重构了针对原始资料整理的微观框架,对于新发现的轻量级原始资料而言,人文学者可以同时进行两项工作:①对原始资料进行数字化和外部结构化描述,对应宏观研究框架的数字化进程,其结果可以存储至关系型数据库中,主要用作资料对应存储和归档。②进行数据分析和数据建模,明确原始资料中包含的各类对象及对象间的关系和属性,将其转化为能够辅助人文研究的中间态数据,并以图数据库的形式进行存储和应用。③上述步骤的目标不在于获取宏观研究框架下数字化和文本化阶段完成后的中间成果,而是基于图数据库存储的中间态数据,进一步实现原始资料的语义化描述和语义关联构建;对于中间态数据而言,既可复用已有的元数据框架将其规范化,实现与宏观研究框架的数据整合,也可以通过语义关联构建转化为语义知识图谱,进而实现与现有知识图谱的融合。

3.3 原始资料整理模型与模型交互

由于数据获取和处理方式以及面向的用户均有不同,根据上文重构的数字人文研究框架,本文提出了以图数据库为核心的、面向人文学科原始资料的整理模型(以下简称"整理模型"),该模型与当前数字人文平台常用的、以知识图谱为核心的知识服务模型(以下简称"服务模型")是两个可交互的独立模型。同时,应将本文构建的图数据库与服务模型中专用于数据仓储的图数据库进行区分,本文构建的图数据库可视为宏观知识服务系统中的一个数据转换模块,与面向公众的知识服务目标不同,其作用领域在人文研究阶段,是数字人文平台开展持续性公众知识服务的必要准备。整理模型与服务模型的交互结构见图 3。

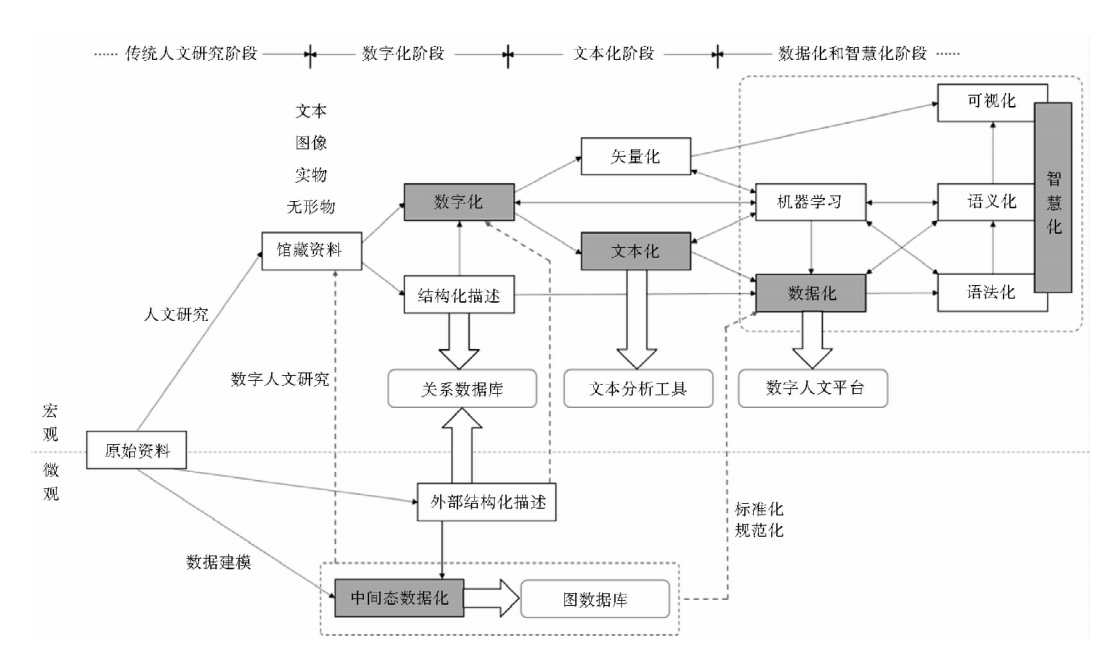


图2　原始资料整理过程与宏观数字人文研究框架的关系

整理模型（以图数据库为核心）
操作要求：非专业知识管理人员
服务范围：学科领域内部用户
应用场景：知识检索、信息推荐、关联构建、……
修正、更新、维护
图数据库：N-E、K-V
数据建模：节点定义、关系定义
原始资料
外部结构化
映射、融合
服务模型（以知识图谱为核心）
操作要求：专业知识管理人员
服务范围：所有用户
知识应用：专家系统、可视化分析、语义检索、推荐系统、问答系统、……
知识计算：路径计算、不一致检测、链接检测、本体推理、规则推理、相似子图推理
知识存储：关系型数据库、图数据库
分布式计算框架：Hadoop、Spark
知识融合：实体匹配、实体对齐、本体融合、本体构建及管理
外部数据库
数据整合
知识抽取：实体抽取、关系抽取、属性抽取
自然语言处理
结构化数据
半结构化数据
非结构化数据

图3　整理模型与服务模型的交互结构

左侧为整理模型，右侧为服务模型，两个模型实现交互的部分共有3处：

第一处，外部结构化交互。指原始资料发现或产生初期，由整理者使用传统手段对资料进行外部结构化描述，并存储到服务模型底层的结构化数据库中，这种对直接感知或观察内容的描述无法对资源内涵进行深度揭示，起到的是一种类似于文献编目的效果，主要意义在于对资源的进行标识。

第二处，中间态数据交互。交互发生在图数据库构建的“数据建模”步骤之后，此时原始资料已经通过分析和建模，被组织成由节点和关系表示的中间态数据，通过中间态数据与服务模型数据进行整合，可以将一部分原始资料纳入现有的知识库中，从而完成对部分原始资料的初次解析。

第三处，语义关联交互。这是图数据库的重要应用场景之一，可以通过图数据库的深度查询功能，辅助用户进行关联构建，从而完成自下而上的本体构建或对服务模型中本体框架的修正。此处交互主要针对创

新性研究价值较高的原始资料，其应用场景对应的是数字人文研究过程，这些研究工作对资料内涵的充分揭示，是原始资料融合到现有知识图谱，并依托数字人文平台面向一般用户开展知识服务的基础。

3 处交互中，外部结构化交互是以传统方法完成的，不再赘述。后文笔者将结合实际，对中间态数据交互和语义关联交互的实现方法展开进一步探讨。

4 考古发掘资料图数据库的构建

图数据库的构建，是实现整理模型与服务模型中间态数据交互的必要条件。本文选取整理任务艰巨，且在人文学科体系中基础性较强的考古发掘资料作为数据对象并构建图数据库，实证数据源自新疆和静察吾呼大型氏族墓葬群的一、四、五号墓地的考古发掘资料[19]。

4.1 考古发掘资料图数据库的构建过程

整理模型所对应的知识服务平台需要涉及到整个系统平台的设计和开发，在保障数据安全性的前提下，相对简易的 B/S 架构(Browser/Server Architecture，浏览器和服务器架构)足以满足有限用户群的中小规模中间态数据的录入、存储和应用需求。本文将着重对 Neo4j 图数据库构建过程和基于 Cypher 的部分业务逻辑的实现方式进行探讨，对于平台架构中的用户交互层面及其相关的数据库访问连接机制不做过多阐述。考古发掘资料图数据库的构建过程主要包括功能分析、数据准备、数据分析、数据建模、图谱生成和知识应用 6 个步骤，其中各个步骤又包含不同的内容，具体如图 4 所示：

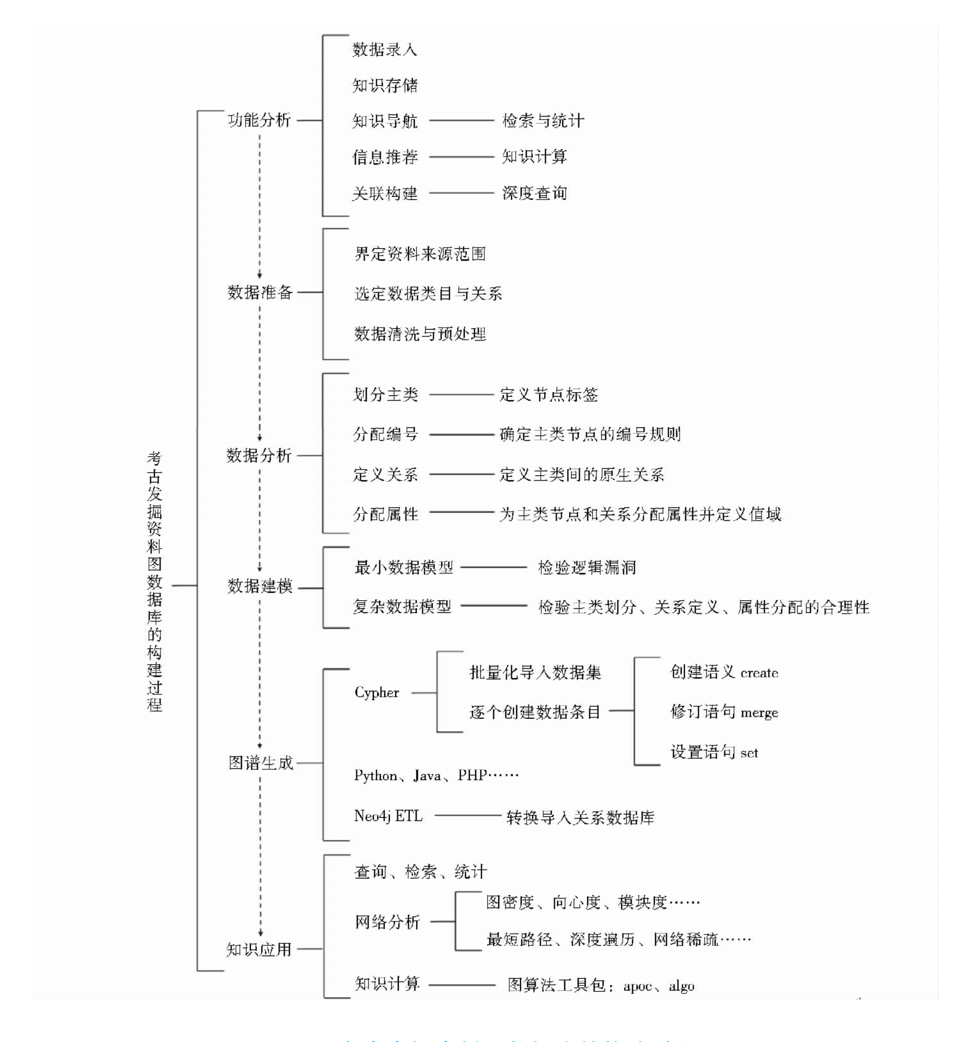


图 4 考古发掘资料图数据库的构建过程

在考古发掘资料图数据库构建过程中，数据分析和数据建模是其中的关键步骤，本节将结合实例着重对这两个方面展开探讨。

4.2 数据分析

数据分析的目的是定义图数据库中需要呈现的节点和关系类型、属性和属性值的类型与定义域。

（1）节点定义。图数据库中的任一节点均需包含一个专属的节点编号（Index:ID）、一个标签（Node:Lable）和若干属性（Property:Keys）。节点的编号需要考虑到各类节点所代表的知识在相关学科体系中的基本分类和层次逻辑，编号本身即可组成基于先验知识的基本知识框架，相较于单纯的顺序编号方法，此种编号规则可做到层次分明、不重不漏且具有一定的可扩展性。节点标签依据数据集子类划分定义，拥有相同标签的节点可视为同类。在非层次网络中，同一类下的各个节点可视为该类实体的实例（Individual）。以考古发掘资料中最常见的墓葬遗迹为例，考古报告中所附"墓葬登记表"所载信息繁简不一，但都会涉及到"遗址""墓形""葬式""墓主信息""出土文物"几类内容，可以将其暂时划定为墓葬发掘资料中的节点主类，在实践中涉及到具体情况时，可以对主类进行相应的调整。

节点属性通常是相关节点的描述性或限定性内容，支持文本型、数值型、向量性等各种形式的数据存储，还能够存储多种格式的图形、动态图形数据和链接，以及 RDF 三元组。节点属性的分配方式并不固定，实践中也需具体问题具体分析。图数据库中节点和关系的属性及属性值还可以通过图计算获得，并进行独立或批量的增删操作。

考古发掘原始资料图数据库的节点定义如表 2 所示：

表 2　考古发掘资料图数据库初始节点定义

节点类型	节点编号	属性 1	属性 2	属性 3	节点标签
	Index:ID	Property:Key1	Property:Key2	Property:Key3	Node:Lable
遗址	$Sid = 1^{\text{类号}} + X^{\text{墓地号}} + Y^{\text{墓葬号}}$ X ∈ [01,99], Y ∈ [001,999]	名称 Name	人数 Bodycount	墓形 Code	Site
	示例:察吾呼五号墓地 M4,型式为 AI,葬 5 具个体 = (101004:Site{Name:"M004_C5", Bodycount:"5", Code:"AI"})				
墓形	$Tid = 2^{\text{类号}} + X^{\text{墓形代号}} + Y^{\text{墓式代号}}$ X ∈ [0001,0009], Y ∈ [1,9]	名称 Name	代号描述 Tdescription	墓形 Code	Tombshape
	示例:AII 式墓 = (200012:Tombshape{Name:"石围石室墓 A 型 II 式", Tdescription:"规则弧腰三角形石围,墓室口距地表较浅,墓室较深,卵石构筑石室,一端开口一端封闭", Code:"AII"})				
葬式	$Bid = 3^{\text{类号}} + X^{\text{二级层次代号}}$ X ∈ [00011,00099]	名称 Name	–	–	Burialform
	示例:侧身屈肢葬 = (300022: Burialform{Name:"侧身屈肢葬"})				
墓主信息(略)	$Oid = 4^{\text{类号}} + X^{\text{二级层次代号}}$ X ∈ [00011,00099]	性别 Gender	年龄段 AgeG	–	Ownerinfo
	示例:成年男性 = (400003: Ownerinfo{Gender:"M", AgeG:"3"})				
文物	$Rid = 5^{\text{类号}} + X^{\text{多级层次代号}}$ X ∈ [10000,99999]	名称 Name	–	–	Relic
	示例:带流杯 AIII = (511230: Relic{Name:"带流杯 AIII"})				

（2）关系定义。图数据库中的初始关系都是直接关系，并未经过进一步的统计、推理和加工。笔者从实证数据集中分离出的初始关系主要有 5 种类型，作为墓葬考古中较具代表性的关系类型，如表 3 所示：

表 3　考古发掘资料图数据库初始关系定义

关系/关系类	源节点	靶节点	属性 - 数量	属性 - 频次	属性 - 比例
Relationship/Type	Src	Dst	Qty	Freq	Pct
墓形为 has_tombshape	遗址 Site	墓形 Tombshape	N	N	N
示例:四号墓地 M160 的墓形为 AII = (104160:Site) - [:has_tombshape] - > (200012:Tombshape)					
葬式为 has_burialform	遗址 Site	葬式 Burialform	Y	N	Y
示例:四号墓地 M160 的葬式包含仰身屈肢葬 = (104160:Site) - [:has_burialform{Qty:"2", Pct:"100"}] - > (300021:Burialform)					
墓主为(略) was_tomb_of	遗址 Site	墓主信息 Ownerinfo	Y	N	Y
示例:四号墓地 M160:男性 2,24 - 30,25 - 30 = (104160:Site) - [:was_tomb_of{Qty:"2", Pct:"100"}] - > (400013:Ownerinfo)					
包含文物 has_relic	遗址 Site	文物 Relic	Y	N	Y

（续表 3）

关系/关系类	源节点	靶节点	属性－数量	属性－频次	属性－比例
示例：四号墓地 M159：带流杯 AIII，2 ＝（104159：Site）－［：has_relic｛Qty："2"，Pct："50"｝］－＞（511230：Relic）					
早于 earlier_than	遗址 Site	遗址 Site	N	N	N
示例：一号墓地 M213 叠压 M279 ＝（101279：Site）－［：earlier_than］－＞（101213：Site）					

Neo4j 图数据库中的关系必须定义方向，但在图节点遍历和其他检索操作中，关系是默认无向或双向的。上述关系中，有 4 种类间关系，1 种类内关系；类内关系“早于”是唯一指向性关系，其所标识的是部分墓葬遗迹之间的地层顺序，这些相对的层位关系在有序考古地层中可用于判断相关墓葬遗迹的相对年代。

关系的类名通常表征关联类型，两个节点和其间关系的类名通常可以视作一个完整表达的 RDF 三元组。关系属性的分配与节点不同，关系属性较多对应的是统计属性，属性值可以应用于图计算，偶尔也用于存储其他内容。

4.3　数据建模

图数据库的数据模型包括最小数据模型和复杂数据模型两种，前者是有关网络初始结构的理论模型，后者则抽取了示例数据源中的真实素材进行建模。

（1）最小数据模型。考古发掘资料图数据库的最小数据模型如图 5 所示，该模型展示了上述 5 个节点类和 5 种初始关系类。图中节点框的框头代表类名，框体代表属性和属性值的数据类型；关系标签中也展示了关系类名、关系属性和属性值的数据类型。

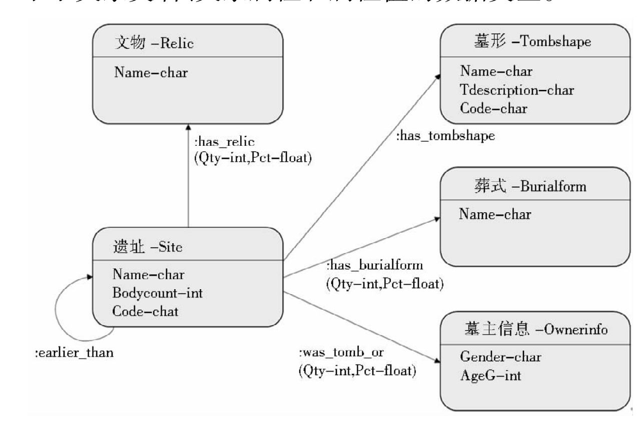


图 5　考古发掘资料图数据库最小数据模型

（2）复杂数据模型。复杂数据模型构建选取本文数据源中的两座具有叠压关系的墓葬进行示例，相关的原始素材为：

素材 1：察吾呼墓地 M052_C4 和 M233_C4 在原报告墓葬登记表中的信息，如图 6 所示：

附表　察吾呼墓地墓葬登记表

表释：♀仰身直肢。♀俯身直肢。♀仰身上屈肢。♀俯身屈肢。♀仰身左屈肢。♀仰身右屈肢。♀俯身右屈肢。♀俯身左屈肢。♀侧身右屈肢。♀侧身左屈肢。♂男　♀女。性别、年龄其中一项不明者用一个×表示，两项均不明者用××表示。如A♀♂×，表示为A个体，仰身左屈肢，男性，年龄不明；A♀××，表示为A个体，仰身右屈肢，性别、年龄均不明。年龄一项省略“岁”字。（二）表示为“二次葬”。F表示儿童祔葬坑。“随葬品”栏中之AI表示A型I式，余同。阿拉伯数字表示件数，不注者表示1件。尺寸单位：米。

附表二　四号墓地墓葬登记表

墓号	型式	方向	石围 长×宽	墓深	盖板或盖木	墓室 长×宽×深	葬具	人数	性别、年龄及葬俗、葬式	随葬品	分期	备注
……												
52	AIII	355°	长方形 5.63×3.10	1.33-1.62	石	2.06×1.50×1.00	无	8	A-F为成人头骨，集中在墓室北端。墓室中有盆、肢、椎、肩胛骨等，归属不明。另有属于两个幼儿的髋骨H、G，性别不明	带流杯AIV3，陶纺轮II，豆把；铜刀CIII，铜锥III，铜管；骨纺轮	三	有墓门。墓门外侧有陶器、羊肋及人的椎骨
……												
233	AII	2°	弧腰三角形残	2.20	无	2.24×0.90×0.60	无	5	A♀♀45±；B（二）头、肢♂40-45；C（二）头、盆、下肢♂25-30；D（二）头、股骨♂×；E♀♀×	带流杯AII、AIII3，勺杯AII、BII，碗I，勺杯A，壶I，双耳罐AI，陶纺轮I；铜针	二	有一马头坑

图 6　察吾呼四号墓地 M052 和 M233 的墓葬登记信息

素材 2：M052_C4 和 M233_C4 的地层叠压信息，如“叠压关系……M52→M233”。

素材中的有效信息可整理出 21 个节点，20 条初始关系，模型表达见图 7。示例中省略了墓主信息及其关系、石围和墓葬规模等细节信息，在实际应用中，这些信息可作为独立节点或节点属性录入图数据库。并对示例中残缺、分类不明确和描述不规范的文物如“豆把”“羊肋”等进行筛除。

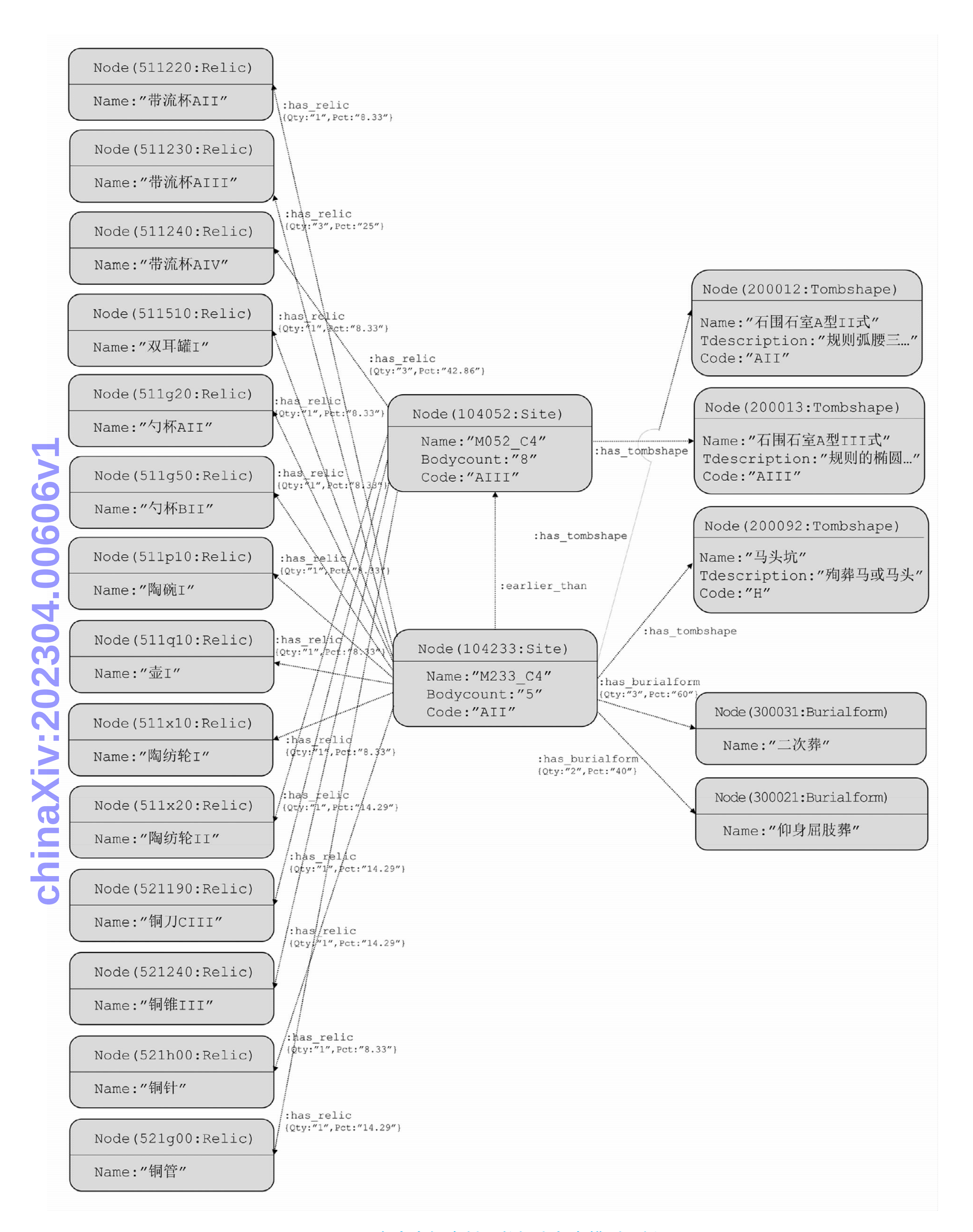


图 7　考古发掘资料图数据库复杂模型示例

通过最小模型图和复杂模型图可以看出本文拟构建的考古发掘资料图数据库结构完整、逻辑清晰，已满足真实数据的导入条件。将真实数据导入图数据库，并去除孤立节点，即可生成新疆察吾呼墓地考古发掘资料的初始图数据库，其中包含 606 个节点，2 739 条关系。这些由图数据库存储的数据资源即可视为原始

资料所转换成的中间态数据，即可以通过数据迁移工具向其他图数据库和关系型数据库共享数据。

5 考古发掘资料图数据库的语义关联

整理模型和服务模型的第三处交互是关联交互，其目标是实现深层次的语义交互，应用图数据库的相关功能，促进考古工作者在对发掘资料研究过程中的知识发现，进而通过知识融合作用于现有的考古学知识服务平台的知识图谱中。其实现主要依靠图数据库的深度查询功能，考古工作者可以由此对原始资料中的一些深层次关联进行挖掘和构建。

5.1 考古发掘资料图数据库的深度查询与关联构建

初始图数据库所录入的都是底层知识节点及初始关联，从关联深度上看，这些关联都属于一度关联。在各人文学科中，均会存在一些具有实际意义的深度关联。以考古学为例，每一地层或遗迹单位（如一座墓葬、窖穴、房基等）中包含的各种遗物所构成的关系被称为文物间的共存关系（Coexistence relationship）[20]，其有助于研究者从整体的文物集合中分离出具有实际意义的固定器物组合方式，发现其中规律，进而能够据此展开年代学分段、文化类型判断等更加细致的研究工作。在此以文物共存关系的构建为例展示考古发掘图数据库的深度查询与关联构建功能。

共存关系是“文物”节点类内的一种二度关联，在察吾呼墓地考古发掘资料构建的初始图数据库中，共存关系的中间节点是“遗址”节点，倘若两种不同类型的文物在同一遗址出土，则视为二者共存，其频次即同时出土二者的遗址个数。以“带流杯 AII（Rid：511220）”和“勺杯 AII（Rid：511g20）”为例，首先，查询二者之间的所有二度关联（见图 8－左），可知同时出土了“带流杯 AII”和“勺杯 AII”的墓葬遗址共计 25 座，统计这些“遗址”节点的数量记为 Sc；其次，在节点“带流杯 AII”和“勺杯 AII”之间建立“共存”（coexistence_with）关系，并将 Sc 值写入关系的频次属性，即可完成两种文物之间的共存关系创建（见图 8－右）。





图 8 带流杯 AII 与勺杯 AII 的共存关系创建结果比对示例（左：创建前，右：创建后）

遍历整个图数据库进行深度查询，共发现并创建了 3 804 种共存关系，以共存关系频次为其赋值的 Cypher 批处理代码如表 4 所示：

表 4 Cypher 批处理代码

```
match (R1:Relic)——(S:Site)——(R2:Relic)
with R1, R2,count(distinct S) as Sc
merge(R1) < -[r:coexistence_with]- >(R2)
setr. Freq = Sc
```

5.2 考古发掘资料图数据库的语义关联示例

图数据库的深度查询与关联构建功能可以将隐藏在扁平的数据网络中、有意义的知识关联提取并展现出来，通过对节点间二度、三度甚至更深度关系的逐层构建，可以在分离实例的情况下保证知识网络的架构完整，以实现更稳定的知识存储并满足更高层次的知识服务需求。此外，广泛应用于知识服务的知识图谱是一种依托本体进行知识组织的语义网，在原始资料中，有可能发现在现有知识图谱中所不具备或未被关注的语义关联，这些语义关联往往正是知识发现的主要目标。以深度查询与关联构建为基础，对语义关联的梳理可视为一种自下而上的本体构建过程，这种基

于图数据库的深层关联创建，可以有效增加领域本体的灵活性和适用性，对于知识体系碎片化且基础性和可变性较强的数字人文知识库建设的作用不言而喻。

以5.1部分实现的"文物"间共存关系为基础，笔者以察吾呼四号墓地M089和M156为例，展示了作为数据层的图数据库和知识图谱之间的连接结构，并从中体现基于图数据库构建的二度乃至深度关联在知识图谱结构中的位置，见图9。

依托考古发掘资料图数据库构建的语义关联可以与服务模型中的知识图谱进行知识融合，已有学者探讨过以图数据存储语义关系的问题[21]，以及RDF与图数据库K-V结构的关联转换问题，实际上，从数据中挖掘和创造关联并用于本体的构建和修正过程，在技术上可以与相关研究互相借鉴。考古发掘资料图数据库语义关联的实现，有助于将考古工作者整理和研究后的原始资料高效、及时、有效地融入现有知识服务平台中，使原始资料实现真正的数据化，加速人文研究成果向可面向公众提供服务的知识产品转化。

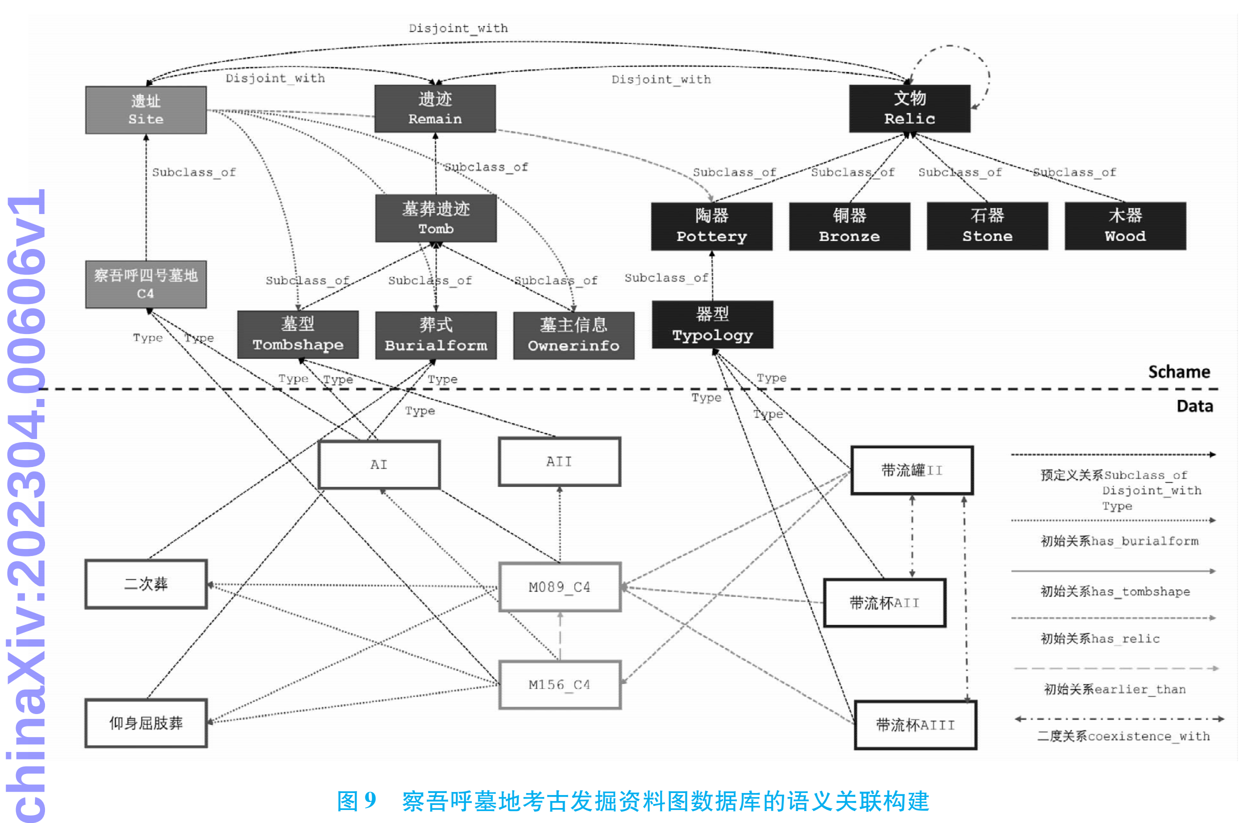


图9 察吾呼墓地考古发掘资料图数据库的语义关联构建

6 结语

首先，本文在明确当前考古发掘资料整理中所存在的现实问题，并对国内外相关的实践项目和研究成果进行综述总结的基础上，结合实际案例，对以考古发掘资料为代表的人文学科原始资料的特征进行解析，提出了将原始资料转化为中间态数据的观点。其次，选取Neo4j图数据库作为原始资料整理的工具，并基于数字人文宏观研究框架重构了原始资料整理的过程框架，结合主流的以知识图谱为核心的数字人文知识服务模型提出了可交互的面向数字人文原始资料整理的图数据库模型，并对两种模型之间的交互形式进行分析。再次，以新疆和静察吾呼墓地的考古发掘资料为例，详细探讨了考古发掘资料图数据库的构建过程，提出了相应的数据分析和数据建模方法，实现原始资料的中间态数据转化。最后，以文物间的共存关系为例，以遍历式深度查询与关联构建技术实现了考古发掘资料图数据库的语义关联构建，为中间态数据的数据化转换以及进一步与知识图谱的知识融合提供思路。后续可以结合自然语言处理、融合情境的相似度计算以及更多的图算法进一步对该方法的应用性功能进行开发，使其在数字人文建设和发展中发挥更大的作用。

参考文献：

[1] 王晓光."数字人文"的产生、发展与前沿[EB/OL].[2021－04－02]. http://blog.sciencenet.cn/home.php? mod=space&uid=67855&do=blog&id=275758.

[2] 林泽斐.英国数字人文项目研究热点分析——基于DHCommons项目数据库的实证研究[J].情报资料工作，2018(1):97

-104.
[3] 刘炜,谢蓉,张磊,等. 面向人文研究的国家数据基础设施建设[J]. 中国图书馆学报,2016,42(5):29-39.
[4] MAY K, BINDING C, TUDHOPE D. A STAR is born: some emerging semantic technologies for archaeological resources[C]// Proceedings of the 36th conference on computer applications and quantitative methods in archaeology. Budapest: CAA, 2008: 402-407.
[5] BINDING C, MAY K, SOUZA R, et al. Semantic technologies for archaeology resources: eesults from the star project[C]// Proceedings of the 38th conference on computer applications and quantitative methods in archaeology. Granada:CAA, 2010: 555-561.
[6] Integrated archaeological database[DB/OL]. [2021-04-02]. http://iadb.org.uk/.
[7] BATTENFELD I, BECKMANN I, SCHULTZE J, et al. Unifying archaeological databases using triples[C]// 4th International conference on cooperation and promotion of information resources in science and technology. Beijing:IEEE, 2009: 281-284.
[8] OCHRE Data Service[DB/OL]. [2021-04-02]. https://voices.uchicago.edu/ochre/.
[9] 张双羽. 考古数据挖掘研究与 E-Arch 考古信息系统优化[D]. 北京:清华大学, 2012.
[10] 祝帆帆,高劲松,梁艳琪. 馆藏文物资源关联数据的创建与发布——以中国十大绘画为例[J]. 图书馆理论与实践,2018(4):96-101.
[11] 郭金龙,许鑫. 数字人文中的文本挖掘研究[J]. 大学图书馆学报,2012,30(3):11-18.
[12] 王丽丽,朱小梅. 古籍钤印元数据著录规范设计与应用研究[J]. 图书馆,2020(1):106-111.
[13] 周耀林,赵跃,孙晶琼. 非物质文化遗产信息资源组织与检索研究路径——基于本体方法的考察与设计[J]. 情报杂志,2017,36(8):166-174.
[14] 周莉娜,洪亮,高子阳. 唐诗知识图谱的构建及其智能知识服务设计[J]. 图书情报工作,2019,63(2):24-33.
[15] 陈涛,刘炜,单蓉蓉,等. 知识图谱在数字人文中的应用研究[J]. 中国图书馆学报,2019,45(6):34-49.
[16] 曾蕾,王晓光,范炜. 图档博领域的智慧数据及其在数字人文研究中的角色[J]. 中国图书馆学报,2018,44(1):17-34.
[17] 孔令远, 陈永清. 江苏邳州市九女墩三号墩的发掘[J]. 考古, 2002(5):19-30.
[18] 马永强, 程卫. 江苏邳州九女墩三号墩出土铜盘铭文考释[J]. 文物, 2019(10):79-81.
[19] 新疆文物考古研究所. 新疆察吾呼——大型氏族墓地发掘报告[M]. 北京:东方出版社, 1999.
[20] 张之恒. 中国考古学通论[M]. 南京:南京大学出版社, 2009.
[21] 王红,张青青,蔡伟伟,等. 基于 Neo4j 的领域本体存储方法研究[J]. 计算机应用研究, 2017,34(8):2404-2407.

作者贡献说明:
高劲松:提出研究思路,设计研究方案,论文修改;
韩牧哲:数据采集与实验,论文撰写与修改。

Research on Semantic Association Construction of the Graph Database on Archaeological Excavation Resources

Gao Jinsong　Han Muzhe
School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] Aiming at the current problems on the collation of primary sources, this paper proposes a method which can realize the datalization and semantic association of archaeological excavation resources, so as to help archaeologists avoid inefficient processes. [**Method/process**] Firstly, after analyzing the characteristics of the primary sources of humanities with example, the process and method of datalization of primary sources was designed; Subsequently, the graph database on archaeological excavation resources was constructed based on the data of Xinjiang Hejing Chawuhu Cemeteries which has been selected as the empirical data source of this paper; Later, we succeed in the association construction of the coexistence relationships between relics, the semantic associations of the graph database were finally realized. [**Result/conclusion**] The construction of the graph database on archaeological excavation resources and its semantic associations could provide a new idea for the datalization of archaeological excavation resources, and has promotional value and practical significance in the field of Digital Humanities.

Keywords: archaeological excavation resources　graph database　semantic associations　Chawuhu Cemeteries